DOCUMENT RESUME

ED 069 737 TM 002 192

AUTHOR Larsson, B.
TITLE An Experimental Study of the Efficiency of Human Information Processing.
INSTITUTION School of Education, Malmo (Sweden). Dept. of Educational and Psychological Research.
REPORT NO R-35
PUB DATE Jul 72
NOTE 53p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Bayesian Statistics; *Cognitive Processes; Hypothesis Testing; Information Processing; *Mathematical Models; *Measurement Techniques; *Neurological Organization; Sampling; Statistical Analysis

ABSTRACT

An experimental study of the efficiency of human information processing is based on the Bayesian model for simple hypothesis testing with fixed binomial sampling. Each of 60 subjects is analyzed with separate ANOVAs focusing on two efficiency variables. Sample size and critical value are also analyzed. Subjects show very different utilization of the independent variables diagnosticity, prior probability and loss, both for their choices and their efficiency of the choices. Giving a part of the experiment as a group test generates similar efficiency results. Efficiency does not seem to be related to intelligence. Final comment connects the experiment with the lens model. (Author/LH)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED 067 17

special-topic
bulletin from

DEPARTMENT OF
EDUCATIONAL AND
PSYCHOLOGICAL RESEARCH

SCHOOL OF EDUCATION
MALMÖ, SWEDEN

TM 002 192

didakometry

Larsson, B.:

AN EXPERIMENTAL STUDY OF THE EFFICIENCY OF HUMAN INFORMATION PROCESSING

No. 35 July 1972

ED 069737

AN EXPERIMENTAL STUDY OF THE EFFICIENCY OF HUMAN INFORMATION PROCESSING

Bernt Larsson

This study is based on the Bayesian model for simple hypothesis testing with fixed binomial sampling. Each of 60 subjects is analysed with separate ANOVAs focusing on two efficiency variables. Sample size and critical value are also analysed. Subjects show very different utilization of the independent variables diagnosticity, prior probability and loss, both for their choices and their efficiency of the choices. Giving a part of the experiment as a group test generates similar efficiency results. Efficiency does not seem to be related to intelligence. Final comment connects the experiment with the lens model.

CONTENTS

Introduction

Data collection

The experiment

The design

The dependent variables

The performance

The group testing

Results

Data processing

The choice between R, R_B, E and E_B

The experiment

Group results on E and E_B

Group results on k_c and n

Individual results on E and E_B

Individual results on k_c and n

The group testing

The decision test and the experiment

Reliability

Other comparisons

Intelligence and efficiency

Final comment

The lens model

Results

References

Appendices

Appendix 1: Choices and expected losses of the statistical model.

Appendix 2: The distribution of subjects on faculty, sex and age.

Appendix 3: The intelligence tests.

Appendix 4: The lens model.

Appendix 5: Symbols used frequently.

INTRODUCTION

This study deals with behavioural decisions and has its theoretical anchoring within Bayesian decision theory. (See e.g. DeGroot, 1970 and Pratt, Raiffa & Schlaifer, 1965.) Although it can sometimes be meaningfully used by those preferring orthodox statistics, Bayes's theorem is a central point for Bayesians. It seems therefore natural that a substantial proportion of Bayesian research is directly concerned with this theorem, e.g. in the form of probability revision experiments. Another substantial proportion is interested in choices of actions and different expected utility theories. While Bayes's theorem tells you how to produce new probabilities when new information reaches you, theories of expected utility tell you how to use them for decision making. One proportion of Bayesian research, which takes both points into consideration, has been labelled information seeking experiments.

Such experiments can involve sequential sampling, fixed sampling or both. Sequential sampling provides the experimenter with more information about subjects than fixed sampling does, but it is as a rule more laborious to perform. Also, if one wants to connect behaviour with statistical theories, these are more complex for sequential sampling than for their fixed sampling equivalents or may even be nonexistent. The most used sampling model, sequential or not, is the binomial one. Two others have been used with some frequency, viz. the multinomial and the normal model. When these models are used in information seeking experiments, they almost always are connected with simple hypothesis testing, while more complex hypothesis testing and point estimation are rare.

Like other fixed sampling models for simple hypothesis testing, the binomial model has three determinants. They are the diagnosticity of data, the prior probabilities and the losses, where only the first one is directly related to the binomial model, while the other two are provided by Bayesian decision theory in order to complete it. Diagnosticity can and has been measured in many ways, both by statisticians and behavioural scientists, and is a measure of how much one observation can discriminate between the hypotheses. It is a function of the difference between the parameter values stated by the two hypotheses. The prior probabilities are the probabilities of the hypotheses, prior to sampling. Losses comprise the "economic" outcome of the choice of a hypothesis and the cost of sampling. Information seeking experiments do not often vary all three determinants simultaneously.

The experiment of this report is mainly chosen to illustrate some new dependent variables. It seems then reasonable to select an experiment which is common within a suitable kind of Bayesian research. Therefore, an information seeking experiment varying all three determinants, as described in the last paragraph, was chosen. However, this study concentrates on the consequences of the subject's decisions and not on how he chooses them, which seems to be of overwhelming interest in the reports issued hitherto. It does not mean that choices are neglected here: two choice variables and two consequence variables will be used as dependent variables.

Although Bayesian experiments seldom hinder you from showing a considerable mathematical machinery, I have not felt this to be necessary, or even desirable, so the mathematics are kept at a minimum. This goes also for data presented. As an unusual feature this study presents hardly any tables (some can be found in appendices) but instead presents important data directly in the text. This may irritate some readers, but it has two distinct advantages. It reduces the number of pages and you can read continuously without interrupting yourself by looking at tables, which perhaps contain only some data of interest for you.

The experiment has an "appendix": the group testing which comprises one decision test and ten intelligence tests. The purpose of this addition is to see whether intelligence is related to efficiency of decision making and whether a group test for decision making can give information equivalent to that of the more expensive experiment. Both the experiment and the tests are discussed in the next section. Although one may argue about how to present individual results, perhaps because we are not so used to these as to group results, I hope that nobody regards them as unimportant. I personally find them at least as important as group results and therefore present several individual results. The final comment makes use of Brunswik's lens model, which I think is a beautiful research paradigm, capable of many applications.

The main questions of this study may thus be put in the following way:

1. How is the choice of the number of observations related to diagnosticity, prior probability and loss?
2. How are the hypotheses chosen?
3. How is the efficiency of the choices related to diagnosticity, prior probability and loss?

4. How is efficiency related to intelligence?
5. How are the efficiency results of the experiment related to those of the decision test?
6. How much do the group results mirror the individual results?

DATA COLLECTION

Data have been gathered at two different sessions, labelled the experiment and the group testing. The experiment is factorially designed and refers, in some degree, to realistic decision situations. The group testing involves intelligence tests and a modified third of the experiment, given as a group test and referring to more hypothetical decision situations. Thus, there are possibilities for comparing individual behaviour in hypothetical and in less hypothetical decision situations and connecting this behaviour with intelligence. Several dependent variables will be used and they also comprise comparisons with optimal behaviour.

The experiment

The experiment uses the statistical model for simple hypotheses testing with binomial sampling. Every situation can be described in the following way: There is an infinite set H with two kinds of element, H_0 and H_1. These elements are in turn infinite sets with elements x which are either 0 or 1 and constitute the observations. The experimenter draws randomly, with probabilities $P(H_i)$, an element from H and from this element the subject draws randonly n observations. The observations are independently and identically distributed with $P(x_j = 1/H_i) = p_i$ $(p_0 < p_1)$ so that $k = \sum_1^n x_j$ is binomially distributed. In common statistical language H_i is called "hypothesis" and $P(H_i)$ "prior probability".

The subject must make two decisions: a choice of n and a choice of H_i. The latter could be guided by the outcome of the observations. A wrong choice of H_i implies a monetary loss c_i, while a correct choice gives zero loss, and every observation must be paid with one unit of the c-scale. The commonest definition of optimal choice of H_i will give an expected loss $L = \min_i (c_i P(H_{1-i}/k,n))$. This so-called Bayes strategy means that H_0 is chosen if $k \leq k_B$ and otherwise H_1 is chosen. The Bayes value k_B is calculated from the equation $c_0 P(H_1/k,n) = c_1 P(H_0/k,n)$. Finally, the optimal choice of n is such that $R_o = \min_n (L + n)$ is obtained. This total expected loss thus refers to non-sequential sampling and will be of particular interest in this study.

The design

A situation is fully characterized by the parameters p_0, p_1, $P(H_0)$, c_0 and c_1 and if an experimenter uses the above model there is often interest to include some of these parameters as independent variables. The most

frequent situations used in Bayesian experiments with the binomial model have $p_0 + p_1 = 1$, $P(H_0) = 0.5$ and $c_0 = c_1$. Less often the experimenter also varies the prior probability or lets $c_0 \neq c_1$, and in some rare cases still greater variation of the parameters is constructed.

To get an idea of the model I have analysed 525 parameter combinations on a computer. The combinations analysed have the following values: $p_0 + p_1 = 0.6\ (0.2)\ 1.4$, $d = p_1 - p_0 = 0.1\ (0.1)\ 0.3$, $P(H_0) = 0.2\ (0.1)\ 0.8$ and $(c_0, c_1) = (150, 600), (200, 400), (300, 300), (400, 200)$ and $(600, 150)$. Among other things, the computer calculated k_B, n_o (the optimal number of observations) and R_o for every situation. Some of the results are presented in Larsson (1970). As there are greater differences in n_o and R_o between d values for constant $p_0 + p_1$ than vice versa, $p_0 + p_1 = 1.0$ was chosen because of greater simplicity, thus eliminating 420 situations. All three d values were included in the experiment though I was doubtful about $d = 0.1$ as most of the R curves (as functions of n) are here very flat around R_o, which implies poor discrimination in R even for rather great variation in n. But 105 situations were too many for an experiment and in the first place I skipped all combinations with $P(H_0) = 0.2$ and 0.8 and $(c_0, c_1) =$ $= (150, 600)$ and $(600, 150)$ because these situations were considered extreme, generating too many situations with $n_o = 0$. As I intended to repeat the experiment there were still too many situations left so, finally, I also took away $P(H_0) = 0.4$ and 0.6.

This leaves you with an experiment where three independent variables (d, $P(H_0)$ and c_0/c_1), which have three levels each, are fully crossed, All independent variables are within-subjects variables so that every subject has the possibility of being compared with the 27 situations. I think that this possibility often generates greater variation in behaviour than the case with between-subjects variables but this is not the cause for the special choice here. The main cause is rather that within-subjects variables give easier comparisons with the group tests where every item is naturally a within-subjects variable. Thus, speaking in the languange of ANOVA, the design of the experiment is 3 x 3 x 3 factorial, all factors being fixed and with repeated measurement. The experiment is given three times to every subject resulting in 81 trials per subject. To avoid order effect the situations are presented in different random sequences to every subject. Appendix 1 shows n_o, k_B and R_o for the 27 different situations.

A H_0 or H_1 was then chosen in accordance with $P(H_0)$ for the 81 trials and for every p value the appropriate number of binomial sequences with n = 1 (1)200 was generated with the aid of a computer. The computer was also used to prepare an extensive table for R values with all combinations of n and critical k values k_c in the range n = 1(1)80 and k_c = 0(1)n. This table will be used to determine values of certain dependent variables described later. (A small number of combinations with n > 80 also needed to be calculated when it was shown that some subjects made more than 80 observations.)

The dependent variables

As in many other kinds of research, the dependent variable in Bayesian experiments can be classified as a choice variable or as a consequence variable. For instance, when a person answers a multiple-choice item the particular alternative chosen constitutes a choice variable, while the evaluation of the item as a correct or wrong answer defines a consequence variable. Although comparisons between a subject's behaviour with the behaviour of a model is far from unusual in Bayesian experiments, choice variables are nevertheless the commonest kind of dependent variables. We have e. g. the number of observations n, the posterior probability $P(H_0|k,n)$ and the likelihood ratio $P(k|H_0, n)/ P(k|H_1, n)$. Concerning the consequence variables used, one may mention the accuracy ratio and different kinds of scoring rules for probability assessments: see Slovic & Lichtenstein (1971) and Staël von Holstein (1970), respectively. A consequence variable often refers to a model: it is a function of two results of a choice variable, the subject's result and the result according to the model. (This is not necessary, the consequence variable can be used to compare two subjects, a subject with a group, etc.)

This study will concentrate on consequence variables but it also contains two choice variables. These are k_c and n. The subject decides to make n observations and selects a critical value k_c such that he chooses H_0 if $k \leq k_c$ and H_1 otherwise. According to the statistical model, the corresponding optimal choices will be denoted k_B and n_o. The consequence variables have to do with losses. However, the actual loss in a situation, which is n, c_0 + n or c_1 + n, will not be used. We will instead use the expected loss R which is a function of k_c and n only (for a given situation). The expectation is over samples: for fixed k_c and n, R is the

arithmetic mean of the actual loss when sampling is repeated an infinite number of times. Coupled with R we define the efficiency $E = R_o/R$, which is examined more closely in Larsson (1970). Due to the definition of R_o the range of E is $0 \leq E \leq 1$. Along with R and E we shall also define R_B and E_B. R_B (E_B) is R (E) corrected for deviation of k_c from k_B, that is R_B (E_B) differs from R_o (1) only to the extent which n is nonoptimal.

Summing up, we have $R = R(k_c, n)$, $R_B = R(k_B, n)$, $R_o = R(k_B, n_o)$, $E = R_o/R$ and $E_B = R_o/R_B$. For a certain situation, k_B is a known linear function of n, but k_c is not in general a known function of n, which means that a construction of $R(k_c, n_o)$ cannot be done in the same way as for R_B. However, R and R_B (E and E_B) will be sufficient, I hope, to give an idea about the partial effects of non-optimal k_c and n. All four consequence variables (R, R_B, E and E_B) will initially be analysed, but only E and E_B will be used throughout as a result of this analysis.

The performance

The experiment was carried through by six persons working at the Department of Educational and Psychological Research, School of Education in Malmö. Every experimenter provided ten subjects. The choice was restricted to subjects who were studying, or had studied on a univeristy level, were not married to the experimenter and had no difficulties in understanding Swedish. The distribution of the sixty subjects as to faculty, sex and age is shown in appendix 2. (The categories are those used later: "Humanities" include one divinity student and two medical students while "Natural sciences" includes four students of technology.) The subjects cannot be regarded as a random sample from a population containing academic persons, such as students in Sweden, nor was it intended to be. Discussion of sample, population and so-called significance tests will be taken up later in connection with the presentation of the results.

After an introduction of the experimental conditions and training of the experimenters, the experiment was performed during three weeks. The experiment was run individually and lasted about 150 minutes per subject. The experimenter introduced the experiment to the subject with the help of written instructions and five training trials. The unit of the c-scale, which equals the cost of one observation, was fixed to 0.1 Swedish crowns. The hypotheses were visualized as two bags, A and B, containing an enormous number of cards, which were either marked with 0 or 1. The

proportions of cards marked with 1 for the two bags were given in writing for each problem. The subject was told that the experimenter had randomly chosen one bag out of many bags, where the proportion of A bags was a certain number, given in writing to the subject for each problem. Then the losses and the observation cost were explained to the subject and they were also given in writing. It was pointed out that many observations made a loss improbable but gave a great observation cost, while few observations gave hardly any observation cost but made a loss quite probable: the subject should consider a balance between these two factors when making observations. The possible outcomes of a certain number of observations was explained. It was said that a great number of cards marked with 1 indicated that the experimenter had chosen a B bag, available to the subject for sampling. On the other hand, a small number of such cards pointed to an A bag. The subject had to decide for a cut-off point: which was the largest number of cards, marked with 1, for which he preferred to guess on A? It was also said that, if he thought so, he could make zero observations and just choose a hypothesis. When he had chosen n and k_c the experimenter told him the outcome k from the simulated binomial sequences. He then wrote down the hypothesis that he chose (as a confirmation) and an estimate of the posterior probability (not used in this report). As we have no interest in learning in this study, no feedback was given to the subject whether he had chosen the correct hypothesis or not. The subject was not paid per hour but had a fixed amount of money from which he had to pay his losses. The subject was told that he could keep the amount left when the experiment was over, and he was also informed what this amount could be at most. This was done to motivate him, but the truth is that the amount left was transformed so that he got something between zero and eighty Swedish crowns, depending on how well he succeeded in relation to other subjects. (The arithmetic mean of this amount corresponded to ten crowns per hour.)

The group testing

The group testing was held within a month after the subject had taken the experiment. It lasted about five hours and comprised eleven tests. One of the tests presents the 27 situations of the experiment in modified form. The modifications are the following: the situations are given in the same

sequence to all subjects, the outcome k is unknown to the subject, he is paid per hour (and does not pay any loss), and the instruction and the test form are therefore somewhat changed. This test will in the following be called the decision test and has the same dependent variables as those described for the experiment.

The other ten tests are proposed to measure some aspects of intelligence. They are selected from a larger pool of tests given to students doing their last term in the "gymnasium". (Students passing this school form qualify themselves for university studies at an age which is usually 19.) The results of this testing is reported in Holmquist (1967). I selected tests which seem to have a tolerable reliability, which do not show any bottom or ceiling effect, and measure several aspects of intelligence. From factor analyses reported in the above paper the selected tests seem to measure (for these students) verbal understanding, verbal fluency, inductive reasoning, spatial ability and perception, two tests for each factor. The intelligence tests are listed in appendix 3 and will be more closely described when results are discussed.

Of the 60 subjects in the experiments only 56 completed the group testing. Three persons were ill and one person left the testing when the last test was given.

RESULTS

After some comments on the statistical treatment of data, there is first a discussion of the choice between different dependent variables. The results of the experiment which are the main points of this report, are then presented in four parts; the divisions are group contra individual results and consequence contra choice variables. The results of the group testing are partly used for a comparison between the decision test and the experiment and partly for a comparison between decision results and the results of the intelligence tests. The section concerning the group testing also comprises discussions of reliability.

Data processing

The statistical treatment of data is based on linear models. Univariate as well as multivariate analysis is used. The attack is wholly descriptive, even if I use methods which by tradition involve inference. This means that the reader cannot find one single probability referring to a significant result in the text. There are several reasons for this. The most important one is that it is very difficult to describe a population of persons to which my sample of subjects can refer. The sixty subjects cannot be regarded as a random sample. Although it is not uncommon in the behavioural sciences to make statistical inferences based on non-random samples I prefer not to do so. However, I will not deny that the results of such samples still contain some possibilities of making generalizations. Such things can also be found in this report, at least as hypotheses, but I find it meaningless to present "exact" significance levels. The generalizations are, by the way, not confined to samples of subjects. We may also have samples of situations and samples of actions, but statistical theory is poorly equipped for this kind of inference.

The second reason concerns the assumption of (multivariate) normal distribution. A good many of the distributions of this study cannot be regarded as normally distributed, some are very different from this bell-shaped "ideal". The talk of robustness, which Bradley (1968, 2.3) has named the myth of robustness, is hardly applicable here, due to the severe deviations from normality. (Also, statisticians have very diverse opinions on this matter.) Non-parametric statistics has not attracted me, because I miss either suitable tests or suitable programs for my purposes.

The third reason has to do with the statistical treatment of separate individuals, where there will be trouble with the assumption of independent observations. Although the situations are randomised for each subject it is not easy to decide whether observations are independent or not between repetitions, which they should be if you want to use ANOVA for inferential purposes. (An interesting question here would also be the problem of generalization: to what behaviour population could you infer from observations of a single person?) As a fourth reason I can add that a significant result has in itself little importance concerning ANOVA for the total group of subjects, because even a very small effect produces a significant F ratio due to the large number of observations.

The elements of the descriptive data presented are arithmetic means, standard deviations and product-moment correlations. Group results and individual results of the experiment are mainly based on ANOVA, the design of which has also been used when discussing standard deviations and correlations. ANOVA of the group results is based on a 3 x 3 x3 x 3 x N factorial design for the total group and subgroups according to sex, age and education. ANOVA of the individual results is based on a 3 x 3 x 3 x 3 factorial design, with one ANOVA for every subject. The basic characteristics of the results here are relevant means and Hays' ω^2, which is explained later. The discussion of the individual ANOVA results has also been supported by a method which identifies outliers. ANOVA has not been used for k_c, because this quality is dependent on the choice of n and is non-numerical when n is zero. I have instead analysed it concerning linear relation to n, both for each situation and for each subject.

No ANOVA has been performed for the decision test but the design is used in a subjective way when comparing it with the experiment. This section comprises the consequence variables only. Besides discussions based on single means, standard deviations and correlations some information comes from canonical correlation analysis and factor analysis. However, neither of them is very convenient: the canonical analysis contains too many variables in relation to the number of subjects and one cannot restrict the weight vectors by suitable hypotheses, the kind of factor analysis available does not give a direct comparison between the decision test and the experiment. These analyses are more convenient

for the comparison between the intelligence tests and the decision results, for which they constitute the main methods. A rather large part of the section concerning the group testing is devoted to discussion of reliability, both for single situations and for sum scores.

It must be underlined that this study contains certain information losses, which does not become more excusable because most studies in the behavioural sciences also suffer from the same "illness". It is understood in most applications of the usual product-moment correlation that if two variables are related then they are linearly related. If not, this correlation can be regarded as a lower bound of the total relation. The product moment correlation is used in this study to discuss certain (minor) results and is a base for reliability discussions, canonical correlation analysis and factor analysis. What a substantial nonlinearity can imply for the result of these analyses is not easy to say. There are methods for checking nonlinearity and my only defense for not having used them is the great amount of extra work they would have involved. However, the main result of this study is free from the above accusations as ANOVA also handles nonlinearity. That nonlinearity is not without importance can be seen from the following example, which refers to the statistical model for the dependent variable n. Here ANOVA shows that the seven effects can predict n perfectly. But only using the three independent variables in an ordinary linear multiple regression analysis must have given (I have not done it) a meager result, since all three variables are nonlinearly related to n.

The choice between R, R_B, E and E_B

If no result will guide the choice, I will prefer E and E_B to R and R_B, because the former variables have absolute scales and involve comparisons with optimal behaviour. The case can also arise that only one of the variables will be chosen. The choice will first of all be based on correlations between the variables, second on reliability and distributions. The statistics are calculated from the whole group of subjects and, as a rule, for every situation, which can mean 108 distributions as we have 27 different situations replicated three times in the experiment and given once as a group test.

The linear correlation between R and E has $-0.998 \leq r \leq -0.761$ with a mean of -0.923 and the correlation between R_B and E_B has $-0.999 \leq r \leq -0.896$ with a mean of -0.969. The latter correlations are, with two

exceptions out of 108, not smaller than the corresponding correlations between R and E (in absolute value). Owing to the high or extremely high correlations one can choose either R or E and either R_B or E_B. The correlation between R and R_B has $-0.236 \leq r \leq 0.965$ with a mean of 0.519 and the correlation between E and E_B varies so that $-0.160 \leq r \leq 0.954$ with mean 0.544. From this it is clear that R (E) and R_B (E_B) cannot be regarded as similar: both have to be used. For situations with low correlations the correction for deviation from the Bayes strategy has far from the same effect on all subjects. However, it is not obvious to me whether to choose R and R_B or E and E_B as the correlation structure is so similar for the two pairs of dependent variables.

The four sets of the 27 different situations may be regarded as a test with 27 items given four times. The square of a multiple correlation, R^2, has been calculated for every item in every set, where the item is regarded as a dependent variable and the other 26 items as independent variables. These correlations can be seen as crude estimates of the item reliabilities (according to classical reliability theory). We have $0.442 \leq R^2 \leq 0.943$ with a mean of 0.777 for R and $0.472 \leq R^2 \leq 0.924$ with a mean of 0.724 for E. We have further $0.479 \leq R^2 \leq 0.988$ with a mean of 0.861 for R_B, while E_B has $0.482 \leq R^2 \leq 0.980$ with mean 0.849. Likewise, the reliabilities of the sums of 27 items do not differ between R and E (between R_B and E_B), but do differ between R and R_B (between E and E_B) as above. Thus, reliability will hardly give any cues whether to choose R and R_B or E and E_B. More will be said about reliability later in another connection.

The distributions of R and R_B are almost all positively skewed, while the distrib tions of E and E_B are positively as well as negatively skewed. If we e. g. define bimodality as a frequency of at least 10 for a class which lies at least three classes away from one or more classes with frequencies of at least 10, R has 2 such cases, E 8 cases, R_B 3 cases and E_B 16 cases. Relative to the standard deviation the class width is somewhat greater for R and R_B than for E and E_B but hardly enough to produce the above differences in the number of bimodalities. If so, E and E_B seem to involve more cases where the subjects are better separated in two groups.

The arithmetic mean m has the following ranges and means: R: $613 \leq m \leq 1396$ with mean 970, R_B: $490 \leq m \leq 1314$ with mean 790. E: $0.513 \leq m \leq 0.911$ with mean 0.724 and E_B: $0.718 \leq m \leq 0.971$ with mean 0.849. We have, of course, $m(R) \geq m(R_B)$ or $m(E) \leq m(E_B)$ for every situation. For the standard deviation s, R has $131 \leq s \leq 764$ with mean 351, R_B has $49 \leq s \leq 314$ with mean 163, E has $0.085 \leq s \leq 0.294$ with mean 0.180 and E_B has $0.048 \leq s \leq 0.236$ with mean 0.136. Here the correction for deviations from the Bayes strategy always gives a reduction of s (R) with $1.0 < s(R)/s(R_B) < 11.1$, but not so for s (E) where $0.8 < s(E)/s(E_B) < 3.8$ with 14 ratios less than 1.0. If anything, this is an advantage for (E, E_B) over (R, R_B) because reduction of s can be assumed to generate fewer differences between subjects.

Summing up, the analysis of the consequence variables has tried to answer two questions. Firstly, do we need all four variables? According to the correlations the answer is no: we need either R and R_B or E and E_B. Secondly, are there any results which point to (R, R_B) or (E, E_B)? There are scarcely such results in the analysis undertaken. We can possibly take the fact that we have cases with $s(E_B) > s(E)$. However, the answer is in principle "no" and for this reason I choose (E, E_B), as mentioned first in this part. Thus, the dependent variables used later in this report will be k_c, n, E and E_B.

The experiment

The treatment of the data builds heavily on the factorial design. Each of the dependent variables E, E_B and n has its own ANOVA, partly for the group of subjects and partly for every individual subject. We have added an ANOVA on n for the results emanating from the statistical model, but not so for E and E_B as all effects will here be trivially zero. The above variables have also been used when a multivariate procedure for identification of outliers is performed. The fourth dependent variable, k_c, is analysed for linear relations with n, both for each of the 27 situations and for every subject (and the statistical model).

Group results on E and E_B

We have primarily analysed the group results with the help of ANOVA as outlined by the experiment. This has been made for the total group and its division according to sex, age and education. Significance tests have been avoided and, instead, descriptive statistics of the different effects

in the form of ω^2 are presented. This index shows the proportion of the total sum of squares which the sum of squares of an effect constitutes, that is $\omega^2_{effect} = SS_{effect}/SS_{total}$. For a closer presentation see e.g. Hays-Winkler (1971, pp. 728-730). The ANOVA gives 31 effects arising from five factors. These are D (different d-values), P (different prior probabilities), C (different cost ratios c_0/c_1), T (different replications) and S (different subjects). Only D, P and C are regarded as proper independent variables of the experiment. Effects containing T but not S inform us about the stability of the group of subjects over replications. Effects containing S but not T inform us about individual differences on several averages. Effects containing both T and S will not be discussed. Likewise, $\omega^2 \leq 0.05$ is considered negligible and I think ω^2 should be at least 0.10 to be of any interest. Of course, this is a wholly subjective statement, but one has to determine a lower boundary and in an exploratory study this boundary could be set rather high.

For the total group the ANOVA of E shows only one substantial effect among the proper independent variables. This is the main effect D for which $\omega^2 = 0.181$. For d = 0.1, 0.2 and 0.3 we have the means 0.812, 0.737 and 0.591, respectively. This result is attributed to different degrees of robustness for different d values. R (k_c, n) is in general steeper around R_o when d = 0.3 than when d = 0.1 for both dimensions k_c and n, which often generates lesser efficiency for d = 0.3 than for d = 0.1, given the same values of $k_c - k_B$ and $n - n_o$. This result is analogous to those of many probability revision experiments, where it is said that greater diagnosticity (d-values) produces greater conservatism (difference between, or other functions of, probability according to Bayes's theorem and estimated probability). Only one further effect is substantial, that of the main effect of S where $\omega^2 = 0.180$. We have $0.428 \leq m \leq 0.893$ with mean 0.714, which I think is quite a good variation for an absolute scale. Values of ω^2 just above 0.05 are found for the interactions SD, SP and SPC.

Compared with E, the ANOVA of E_B for the total group exhibits raised ω^2 values for S and SD and a lower value for D, other things being essentially the same as for E. For D we get $\omega^2 = 0.109$ arising from the means 0.904, 0.865 and 0.778. Comparing these values with the corresponding ones for E, we find that the correction is most bene-

ficial for d = 0.3. Further, the values of E_B - E and 1 - E_B are about the same for every d, meaning that the inefficiency is equally caused by nonoptimal choice of n and k_c. For S we now have ω^2 = 0.270 with 0.644 $\leq$ m $\leq$ 0.966 and mean 0.849. For SD ω^2 = 0.162 which can be illustrated by three D profiles which are most different among themselves: (0.947, 0.929, 0.929), (0.585, 0.754, 0.591) and (0.945, 0.699, 0.449). Thus, relative to the total sum of squares we have a better differentiation of the subjects for E_B than for E. (SS_{total} for E_B is about one half of SS_{total} for E.) No other effects are over 0.05 and, especially, effects containing T but not S are far below the 0.01 level. This is also true for E so that the group does not change in behaviour from replication to replication. Or more exact, their behaviour is such that the consequence of the behaviour is the same from replication to replication.

We may construct an average subject through calculating means over the sixty individual subjects. The ANOVA of this average subject can be deduced from the ANOVA of the total group if all effects containing S are ignored. Doing this, we get values of ω^2 which are small for all but two effects. For E and E_B we have ω^2_D equal to 0.757 and 0.671 and ω^2_{DPC} equal to 0.129 and 0.196, respectively. Thus the efficiency of the average subject is very dependable on different d values.

No computer program was available which could incorporate sex, age and/or education as extra factors in the above design, because the number of cells became too large. I have therefore made ANOVA as before, one for men, one for women, etc., which is a little unsatisfactory as e.g. all effects involving sex cannot be directly evaluated. Anyhow, it seems to me that the new ANOVAs tell approximately the same story as did the ANOVAs of the total group. Thus, I will comment briefly upon the results.

Concerning sex, men have ω^2_D = 0.136 and ω^2_S = 0.206 and women have ω^2_D = 0.244 and ω^2_S = 0.149 for E, while for E_B we have 0.054, 0.270, 0.192 and 0.235, respectively. For E_B we further have ω^2_{SD} = 0.190 for men and ω^2_{SD} = 0.115 for women. Other effects have ω^2 not greater than 0.066 and often are much smaller. Relative to their own sex, men are less affected by different d values than women are and are more differentiated in their means. However, the mean efficiency is about the same for both sexes, being 0.725 for men and 0.701 for women, concerning E, and 0.873 and 0.820, respectively, for E_B. For both sexes E_B - E is greatest for d = 0.3 and at least here we have a pronounced difference:

while men have the same value E_B - E and 1 - E_B (a small positive difference), women's inefficiency is more related to the choice of n than to the choice of k_c, given n: E_B - E - (1 - E_B) is -0.120.

The total group is divided into three age groups, i.e. A1: at least 30 years old, A2: between 25 - 29 (inclusive) and A3: at most 24 years old. For E one finds ω^2_D = 0.204, 0.206 and 0.166 for A1, A2 and A3, respectively, and the corresponding values for ω^2_S are 0.147, 0.163 and 0.184. Three other effects have $0.080 \leq \omega^2 \leq 0.090$ for A1, but we have as a whole no difference between the age groups. The case of E_B has ω^2_D = 0.152, 0.082 and 0.129, ω^2_S = 0.287, 0.282 and 0.233 and ω^2_{SD} = 0.166, 0.127 and 0.188. Again we have the same picture: ω^2_D goes down and ω^2_S and ω^2_{SD} rise, when E is replaced by E_B, although in somewhat different degrees for A1, A2 and A3. No total mean differences between the groups are discovered; E gives 0.662, 0.700 and 0.735 while E_B gives 0.811, 0.844 and 0.861, but the trend is that the younger subjects are a little more efficient. As for the sexes, E_B - E and 1 - E_B grows with increasing d values and E_B - E is in most cases slightly smaller than 1 - E_B. There are two exceptions: for d = 0.2 the A2 group is much more affected by the choice of k_c than by the choice of n, E_B - E - (1 - E_B) being 0.103. For d = 0.3 and A1 the corresponding value is -0.095.

The total group has also been classified as to type of academic study with special regard to mathematics and statistics. The three groups are E1: humanities, E2: social sciences and E3: natural sciences, the distribution of which was given in appendix 2. One may assume that good knowledge of mathematics and statistics will produce better efficiency than little such knowledge. This hypothesis has been examined before, see e.g. Kogan & Wallach (1964). Although there are overlaps, it is reasonable to suppose that E1 has the least mean knowledge, E3 the greatest mean, while E2 will take a middle position. For E, ω^2_D = 0.244, 0.186 and 0.110 for E1, E2 and E3, respectively. We have further ω^2_S = 0.142, 0.187 and 0.119. For E_B, ω^2_D = 0.218, 0.084 and 0.074, ω^2_S = 0.190, 0.308 and 0.107 and ω^2_{SD} = 0.150, 0.150 and 0.177. With the exception of E3 for factor S the same picture reappears: ω^2_D becomes smaller and ω^2_S and ω^2_{SD} becomes greater. However, there are greater numerical differences here than for the other classifications. For instance, E1 is much more affected by different d values than E3 is and E2 has more differentiated individual means than E3 has. Other effects are small,

although E3 has some minor ones, e.g. $\omega^2_{DPC} = 0.062$ and 0.090 for respective E and E_B. The total means are for E 0.713, 0.695 and 0.773 and 0.824, 0.846 and 0.892 for E_B. Thus, the hypothesis about knowledge of mathematics and statistics is in line with the above means, but the differences in these seem to be too small for a real confirmation of the hypothesis. Again we have increasing values of $E_B - E$ and $1 - E_B$ for increasing d values for all three groups. For d = 0.2 and 0.3 E1 has $1 - E_B > E_B - E$, while for the other groups the choice of k_c and n produce about the same inefficiency. Notice that there is a certain correspondence between sex and education: the eleven students of natural sciences consist of ten men, while the sixteen students of humanities have only four men. In fact E1 and women have many similar results on ANOVA and E3 and men have some corresponding results.

Group results on k_c and n

The main results come from ANOVA on n and the linear relation between k_c and n, both analyses for the total group only. The ANOVA shows only one substantial effect, that of S which has an ω^2 of 0.537 and this refers to means between 0.0 and 91.0 with a total mean of 21.1. No other effects give ω^2 greater than 0.05. The sum of ω^2 for the proper independent variables D, P and C is 0.022 and the corresponding sum for effects containing T but not S is 0.003. We can certainly say that, relative to the variations between the subjects, the choice of n is constant over replications and scarcely dependent on the different situation parameters.

Looking at the average subject, whose ANOVA contains the above sums of squares, which do not contain S, we find three ω^2 of some size. These are $\omega^2_D = 0.134$, $\omega^2_P = 0.458$ and $\omega^2_{PC} = 0.183$. For D the means are 23.2, 20.6 and 19.5 and for P we have 19.1, 25.1 and 19.0, where the first mean of each effect corresponds to the lowest level, and so on. Concerning PC we have (21.9, 17.8, 17.6), (25.4, 24.7, 25.2) and (17.0, 16.9, 23.0) for the simple C effects of $P(H_0) = 0.3$, 0.5 and 0.7, respectively. Comparing with the statistical model, the total means are almost the same: 20.6 due to the model and 21.1 for the average subject. The model produces three ω^2 above 0.100, i.e. $\omega^2_D = 0.676$, $\omega^2_{PC} = 0.142$ and $\omega^2_{DPC} = 0.155$. Also, the sum of ω^2 of effects containing T is zero while the average subject gives a sum of 0.123.

The greatest difference between the model and the average subject comes from the choice of n for d = 0.1, where the model has a mean of 7.4. This arises from the fact that for the asymmetric situations, when d = 0.1, the impact of the observations is so slow that it is optimal to choose a hypothesis without paying any observations: $R(k_B, n) \geq R(k_B, 0)$. That subjects disagree with the model in this way for similar situations has been verified before, see e.g. Larsson (1968) and Snapper & Peterson (1971). However, as most functions R are flat around R_o for d = 0.1 the inefficiency of this disagreement is in general insignificant. As we shall see later, the above "wrong" choices of the average subject are not valid for all individual subjects.

The model and the average subject both behave in the same way for different prior probabilities, although the average subject has a greater variance of the means. (Notice that for an effect i we have $\omega_i^2(a)/\omega_i^2(b) = [SS_i(a)/SS_i(b)]\,[SS_t(b)/SS_t(a)]$, where SS_t is the total sum of squares and a and b denote two persons, etc. If we let a stand for the average subject and b for the model we have the following relation for the main effect P: 0.458/0.014 = 660/150 · 10885/1441 . Thus the great ratio between the ω^2's is dependent on a greater variance of the P means of the average subject and his lesser total variance.)

The PC effects show about the same patterns; the exception is C for $P(H_0) = 0.5$, where the average subject produces a horisontal profile and the model a triangular one. The model, however, has a greater variation than the average subject. Finally, the DPC effect of the model is ordinal, meaning that the different simple interactions of PC show the same pattern but with different spreads e.g. the three profiles of $P(H_0) = 0.3$ are all non-increasing for increasing C levels.

As the Bayes strategy implies that k_B is a linear function of n_o it has been natural to me to analyse k_c as a linear function of n: to what extent and how can we express k_c as A + Bn? For the model all 27 situations give B = 0.5 with $-3.84 \leq A \leq 3.84$ for d = 0.1, $-1.90 \leq A \leq 1.90$ for d = 0.2 and $-1.24 \leq A \leq 1.24$ for d = 0.3. Due to symmetry we have the same A value but with reversed sign when a situation with parameters $(d, c_0/c_1, P(H_0))$ is replaced by a situation with parameters $(d, c_1/c_0, 1 - P(H_0))$.

The linear relation between k_c and n has been analysed for each of the 27 situations with at most 180 cases (60 persons times 3 repetitions) for

every situation. Cases with n equal to zero have been deleted, as k_c has no numerical value for these cases. This means that the number of cases varies between 114 and 169. The linear relation is clear: the correlations between k_c and n is such that $0.787 \leq r \leq 0.995$ with a mean of 0.950. For the A and B values we have $-0.566 \leq A \leq 1.840$ with a mean of 0.587 and $0.233 \leq B \leq 0.488$ with a mean of 0.402. It is thus obvious that persons tend to underestimate k_B, at least when n is great. In fact, there are situations where persons underestimate k_B for all $n > 0$. This behaviour is an important factor when explaining inefficiency: nonoptimal choice of great n values is combined with bad choice of k_c. Considered as a group, these persons have a clear bias against the hypothesis with the smaller p value, at least for great n. Why this is so is difficult to understand. A tentative explanation is that most subjects overestimate the information of a "1" in relation to the information of a "0". Another concerns the instruction given to the subjects: "I choose H_0 if the number of ones is less than or equal to _____." Perhaps we had got the opposite bias if the instruction had been "I choose H_1 if the number of ones is greater than ____." The bias is related to the factors D and P, such that the bias is greater for greater d values and smaller for greater $P(H_0)$ values. (We have B equal to 0.440, 0.407 and 0.360 for D levels and 0.350, 0.412 and 0.444 for P levels.) The relation to P is quite "reasonable", and similar to the statistical model, but the relation to D is harder to suggest explanations for. Anyhow, this relation also generates inefficiency because of the lesser robustness to deviations from k_B for d = 0.3 than for d = 0.1.

Let us again look at an average subject. You may imagine him in the following way: Every subject chooses a n value and a k_c value for every situation and repetition, and the average over subjects and repetitions constitutes the choice of the average subject for a certain situation. This means an average n which is calculated from 180 cases. As some of the k_c values are non-numerical the average k_c value is calculated from A + Bn, where A and B is the average estimated parameters discussed above and n is here the average n value just mentioned. (Strictly speaking, n is the nearest integer to this average n and k_c is then the greatest integer less than or equal to A + Bn.) For most situations both E and E_B are greater for this average subject than for the average E values of the subjects. This was expected since most efficiency curves, as functions of n or k_c and n, are concave. The differences are greater for

d = 0.3 than for d = 0.1, especially for E_B, because the concavity is more pronounced for greater d values. We have $0.481 \leq E \leq 0.991$ with a mean of 0.780 and $0.775 \leq E_B \leq 1.000$ with a mean of 0.954. The inefficiency is almost always little dependent on the choice of n. Thus, this type of group decision will in general improve on the choice of the amount of information but not on how to use it. However, exceptions from this "rule" can be found for certain situations and there are also certain subjects who are more efficient than this average subject (or group decision).

Individual results on E and E_B

For every subject there is an ANOVA with factors D, P, C and T (compare the group ANOVA), both for E and E_B. Effects are considered nontrivial only if ω^2 is greater than 0.100. There is great variation between subjects, showing from zero to five substantial effects in their ANOVAs; about half of them show two effects. The commonest one is D, then comes DPC, just as for the average subject, constructed by collapse of the group ANOVA. While there are subjects with about the same pattern as this average subject, there are also subjects with totally different "styles", e.g. the one with no substantial effect. This does not mean that he behaves like a statistician: the average efficiency can be far from 1.000 and/or his variation, concerning efficiency and therefore his choice of k_c and n, from repetition to repetition may be great. This is in fact the case for the subject with zero effects.

It is almost impossible to go into details of every ANOVA. I have instead selected some ways of description to highlight individual differences. One of these ways concerns the identification of outliers, which has been performed by a multivariate technique based on the Mahalanobis distance. This has been done for every repetition, for E as well as for E_B. The method selects the subjects (if any) who are "too far away" from the group centroid in the 27-dimensional space, which constitutes the space where subjects are represented as points for our case. (See Dixon (1970), pp. 104-112.) The selection, of course, results in a more homogeneous group, as concerns the remaining subjects. They are also better: the subjects deleted are in general inefficient and this is valid for E and E_B. It is not always the same subjects who are selected in the six cases, and those who are may come in different order from case to case. (Let rank order 1 denote that the subject is selected first, and so

on). I have looked at the ANOVA results for the subjects who have the five lowest average ranks, partly for E and partly for E_B. We have six subjects totally, four persons are the same for both dependent variables. One of them has means which correspond to the group means, but the others are far below these levels.

It has been stated earlier that for most situations we have $s(E_B) \leq s(E)$ and that $SS_{total}(E_B)$ is about 50 per cent of $SS_{total}(E)$ for the group ANOVA. The same reduction is, as a rule, also found for the individual ANOVAs. There are subjects whose $SS_{total}(E_B)$ is only 5 per cent of $SS_{total}(E)$, depending on ceiling effects: the correction for deviations from the Bayes strategy makes the efficiency values high. On the other hand there are subjects with no reduction and three of those are among the above-mentioned outliers. One may expect that E_B shows smaller variance than E, since one of the causes for inefficiency has been removed, and this is usually true. But if a subject almost always chooses $k_c = k_B$ or if his k_c choice is very varied there need not be any reduction, on the contrary, there can be an increase of SS_{total}. The three outliers are of both types: one subject has $1 - E_B$ and $E_B - E$ equal to 0.333 and 0.051, while the others have (0.356, 0.183) and (0.335, 0.237) and thus are inefficient when choosing n as well as k_c. While $1 - E_B$ and $E_B - E$ are of the same magnitude for the total group (0.151 and 0.133, respectively), we find great variations among the subjects. All four types are represented: good at both k_c and n choice (example: 0.067 and 0.050), good at k_c and bad at n choice (example: 0.333 and 0.051), bad at k_c and good at n choice (example: 0.086 and 0.279) and bad at both choices (example: 0.335 and 0.237). If we make median splits for $1 - E_B$ and $E_B - E$ the cell frequencies of the fourfold table are 18, 12, 12 and 18 which means a smaller negative correlation than I had expected with regard to the construction of the variables.

The six outliers are all but one worse than the average, as concerns the choice of the number of observations. All of them show about the same ω^2 profile: they have fewer substantial effects and these are lower than average. We have, with results from all individual ANOVAs within parenthesis, for E the mean number of substantial effects equal to 1.00 (2.17) with $0.102 \leq \omega^2 \leq 0.210$ ($0.100 \leq \omega^2 \leq 0.755$) and for E_B the corresponding results are 1.17 (2.02) and $0.101 \leq \omega^2 \leq 0.328$ ($0.101 \leq \omega^2 \leq 0.788$). This implies that every outlier has small differences

between means and/or great variations over repetitions. The first cause is more valid for E and the second one for E_B. (They have about average SS_{total} for E but above the average SS_{total} for E_B.) They tend to act like random number generators, when it concerns the choice of n: sometimes they hit the target and sometimes they are far from the optimal number.

The variation between subjects is not the same from situation to situation. For factor D the greatest differentiation is obtained for d = 0.3 with standard deviations (0.169, 0.180, 0.205) for E and (0.109, 0.125, 0.190) for E_B. This is a reasonable result, as the robustness of efficiency, as to choices of k_c and n, is greater for d = 0.1 than for d = 0.3. Hence, a certain variation of choices causes greater variation for the greater d value. No more systematic effect can be discovered for E_B, but for E there is another effect, which can be seen for P and C and which is very pronounced for PC. With increasing $P(H_0)$ values we have, for increasing c_0/c_1 values, (0.169, 0.182, 0.233). (0.170, 0.155, 0.175) and (0.223, 0.179, 0.143). It is quite evident that asymmetrical situations produce greater differences between subjects than more symmetrical situations. As this is not the case for E_B, the fact must be caused by the choice of k_c. The figures 0.233 and 0.223 refer to the situations where H_0 is both probable and cheap (when wrongly chosen) and where H_0 is both improbable and expensive, respectively. There are obviously more different opinions as to how to choose k_c when both determinants "go in the same direction". Perhaps the smaller variations in the opposite situations are due to some general reasoning like "the two factors will balance each other so I should choose k_c near n/2"?

Individual results on k_c and n

For every subject there is an ANOVA of n like those for E and E_B. We also have an analysis of the linear relation between k_c and n for every subject, including the statistical model. We have again used the method for detection of outliers, as concerns the choice of n. Let us again select the subjects with the five lowest average ranks. It is interesting to notice that three of the outliers from E and E_B reappear (the three subjects which have no reduction of $SS_{total}(E_B)$). Some of the characte-

ristics of the outliers for n are the following, with corresponding results for all individual ANOVAs within parenthesis: they make many observations, $29.9 \leq m(n) \leq 91.0$ ($0.0 \leq m(n) \leq 91.0$), they have the five greatest SS_{total}, they have an average number of $\omega^2 \geq 0.100$ of 1.4 (1.3), but these are small, $0.106 \leq \omega^2 \leq 0.251$ ($0.100 \leq \omega^2 \leq 1.000$). Thus, the outliers make too many observations, have no pronounced strategies for the choice of n and lie below the 20th percentile on both E and E_B; in fact we find the worst subject on each efficiency variable among these outliers.

The individual ANOVAs for n have together about half as many effects with $\omega^2 \geq 0.100$ as the ANOVAs for E and E_B, and they are otherwise distributed. Most common effects are D, P and PC. There are no more subjects with distinct ω^2 profiles here than for E and fewer than for E_B. (7, 6 and 15, respectively, if we define a distinct profile as one with either a sum of the substantial ($\omega^2 \geq 0.100$) effects greater than 0.750 or one which has a single effect greater than 0.600.) The strategies of information purchase as illustrated by the ω^2 profiles (or even the distinct ones) are quite different between the subjects. One subject is most sensitive to D ($\omega^2 = 0.525$), three others concentrate on P ($\omega^2 = 0.659$, 0.779 and 0.794), another one on C ($\omega^2 = 0.901$), while one is totally absorbed by PC ($\omega^2 = 1.000$) and the other subjects more or less have strategies which take into consideration more than one effect. There are ten subjects with no substantial effect at all and hence with no strategy, except a random one. (Another two subjects always make the same number of observations, which implies that their ω^2 values are not defined.) I may also inform the reader that the statistical model produces a strategy, which concentrates on D ($\omega^2 = 0.676$).

The distributions of n are all positively skewed. Surprisingly many subjects have chosen n equal to zero (between 2 and 24 per trial), but at the same time there are almost always choices of at least 100 observations (between 0 and 8 per trial). The standard deviations are therefore of the same magnitudes as those of the means. We have $11.5 \leq s \leq 42.6$ with an average s of 25.8. According to the factorial design there is but one effect of s. For increasing D levels the means of s are 30.9, 24.0 and 20.7, which seems reasonable, for the following reason. If we could plot n as a function of d (average over P and C) for

every subject and many d values, we probably would obtain curves which were unimodal and with n equal to zero when d is zero and one. The position of the maximum n value and the average n of a curve are different from subject to subject. From this we will expect small s values for very low and high d values. In our case we can expect still smaller s values for d greater than 0.3 than the s value for d equal to 0.3. If we had made d smaller than 0.1 we could expect that more and more subjects would ultimately realize the futility of making any observations. Although the curve generated from the statistical model has its maximum n value when d is about 0.2, only a few subjects have the same type of a D profile. More than half of the subjects have profiles which vary less than five units of n. Another 15 subjects have profiles where n decreases for increasing d values.

The linear relation between k_c and n has been analysed for each of the subjects with at most 81 cases for a subject. When n is zero the case has to be deleted as k_c is non-numerical here. The number of cases varies between 0 and 81, but only five subjects have less than 45 cases. The linear relation is more or less evident from subject to subject: we have $-0.022 \leq r \leq 0.997$ with a mean of 0.820. There are 28 subjects with r greater than 0.900 and only ten subjects with r less than 0.700, and five of these values depend on $s^2(n)$ being zero or very close to zero. For the A and B values ($k_c = A + Bn$) we have $-8.094 \leq A \leq 9.500$ and $-0.050 \leq B \leq 0.866$ with means of -0.039 and 0.426, respectively. We have, on the average, the same results here as for the total group: the individual subjects are in general biased against the hypothesis with the smaller p value. However, the differences between subjects are great. We have a few subjects which are biased against the other hypothesis, some subjects are not biased, while some subjects are so biased against the hypothesis with the smaller p value that they always choose k_c smaller than n/2. The standard deviation of k_c, given a particular n value, also varies greatly between subjects: $0.2 \leq s \leq 11.3$, where s stands for the standard deviation of k_c about the regression line. Four of the above mentioned five outliers for n have the four greatest s values and they are all biased against the hypothesis with the smaller p value. Only one of them belongs to the subjects with the five greatest values of $E_B - E$. The latter subjects

have rather great s values but three of them are not biased. Why this is so, cannot be settled by the analysis here. Perhaps these three subjects choose k_c far from k_B for situations which are not robust for deviations from k_B, but I do not know. It can be added that the statistical model gives an r value of 0.921 with a standard deviation about the regression line of 0.9 and that A = -0.500 and B = 0.500. (A is different from zero, because k_B is an integer and this produces a bias.)

The group testing

This part deals with comparisons between the experiment and the decision group test and the intelligence tests. The presentations concern group results only and for the experiment as well as for the decision test the dependent variables are limited to E and E_B. The comparisons use means, standard deviations and correlations. The correlations are further analysed by the use of canonical correlation analysis and factor analysis. Discussions of reliability are also made.

As has been stated before, all 60 subjects did not take the tests. The results of the decision test are based on 57 full records, while the results of the intelligence tests comprise only 56, since another subject had to be deleted. Looking at the results of the experiments, the greatest differences between those deleted and the total group are found for E, as concerns the decision test. (Means of 0.623 and 0.714, respectively.) If we suppose no change of results from the experiment to the decision test, the deletion will cause an increase of the total mean to 0.719, which can be considered negligible. Still lesser effect may be expected for standard deviations. E.g. the standard deviation of the subjects' means of E will, under the above assumption, not change more than 0.001. On the whole I do not think that this five to seven per cent of non-response is anything to worry about: differences of results between the decision test and the experiment is hardly due to differences between the 57 subjects and the 60 subjects.

The decision test and the experiment

Reliability

We will begin with some viewpoints on reliability. Every situation can be regarded as an item and for each repetition of the experiment, as well

as for the decision test, there are possibilities for observing reliability of an item and of the sum score of the 27 items. This can be done in several ways, both with regard to the definition of reliability and the estimation of reliability. We have, in principle, three populations: those of subjects, items and actions. No generalizations will be made as to a population of subjects, since the subjects of this study cannot be regarded as a random sample. Nor will generalizations be made as to a population of items. The definition of this is in general very difficult, but we have the unusual possibility of defining the population unequivocally according to factor D, P and C. However, the 27 situations selected are hardly any random sample from such a population. As a consequence of this the above factors have been regarded as fixed for the ANOVAs. The only population left is difficult to discuss, because it is not obvious how to define a random sample. So, strictly speaking, there are no generalizations for the reliability values, which is in accordance with what has already been stated about the study at large. On the other hand, I think it is reasonable to expect the same kind of results, as have been found here, if you replicated the experiment, even if you chose some other levels of D, P and C and, perhaps, also with other, similar subjects.

When we speak about reliability here, we refer to the classical model, see e.g. Lord & Novick (1968, ch. 3). Let us start with the item reliabilities. Two measures are used: one internal measure (within a set of 27 items) and one measure based on correlations between the sets. The internal index is the squared multiple correlation between an item and the other 26 items. If the number of subjects is very great in relation to the number of items, the squared multiple correlation R^2 is a lower limit - and perhaps a bad one - of the reliability. However, when the number of items approaches the number of subjects, R^2 will approach 1. A common correction for this bias is based on the unbiased estimate of the residual variance, see e.g. Darlington (1968). I believe that the corrected values better reflect facts, because they seem less affected by the relation between the number of items and the number of subjects. (Notice that this correction need not concern inference: the same bias is obtained whether we call our subjects a sample or a population. According to Dempster

(1969, p. 161) "theoretical understanding of this phenomenon of diminishing returns for variables introduced remains imperfect, ...".)

The following squared multiple correlations are obtained (with uncorrected values within parentheses). The decision test has 0.145 (0.542) $\leq R^2 \leq$ 0.763 (0.873) with a mean of 0.504 (0.735) for E and 0.033 (0.482) $\leq R^2 \leq$ 0.875 (0.934) with a mean of 0.633 (0.804) for E_B. The experiment has 0.055 (0.472) $\leq R^2 \leq$ 0.862 (0.924) with a mean of 0.500 (0.721) for E and for E_B we have 0.309 (0.614) $\leq R^2 \leq$ 0.962 (0.980) with a mean of 0.756 (0.864). Two results are obvious: the reliability of E_B is better than that of E and the reliability of the items of the decision test is equally good as for the experiment when it concerns E but lower for E_B (on the average). In spite of the smaller standard deviations of E_B, the reliability is greater here, because there is only one unreliable determinant: the choice of n. However, this is not always so; in 11 cases out of 108 we have the reverse relation. The average item reliability must be regarded as good.

The correlations between replications of the situations can be regarded as (modified) retest correlations. For the experiment the modification consists of the items being presented to the subjects in different random orders. The decision test is so different from the experiment that I hesitate to call the correlations between this test and a repetition of the experiment retest correlations. Yet I give them - they may have interest as lower boundaries. The decision test has $-0.085 \leq r \leq 0.566$ with a mean of 0.267 for E and $-0.051 \leq r \leq 0.606$ with a mean of 0.306 for E_B. The corresponding values of the experiment are $0.255 \leq r \leq 0.824$ with a mean of 0.509 and $0.178 \leq r \leq 0.902$ with a mean of 0.620. The experiment shows the same average (for E) here as the average of R^2, while the r mean of E_B is smaller than the R^2 mean. As was expected the correlations are smaller for the decision test.

It can be added that the reliability values calculated from the group ANOVAs on E and E_B give average item reliabilities of 0.372 and 0.528, respectively. An interesting feature is that the corresponding value for n is 0.705. This can depend on two things. Since the reliability calculated from the ANOVA is an intraclass correlation, it is only equal to the average correlation between repetitions if all 81 variances are equal.

So different deviations from this may give the difference between 0.705 and 0.528, even if correlations between repetitions are, on the average, of the same magnitude for n and E_D (E is not quite comparable here, as it is also dependent on the choice of k_c). The other cause is more credible to me, and that is that n is more reliable than E_B. One indication of this is the high correlations between n and k_c with an average of 0.950. According to the classical theory such a correlation is a lower boundary of the geometric mean of the reliabilities of n and k_c which, by the way, show that also k_c has a high reliability. I think it is reasonable to expect higher reliability on a choice variable than on a consequence variable. The latter is a transformation of the former, which sometimes (not here) involves unreliability in itself, e.g. at subjective judgments of different kinds. But even when the transformation can be mathematically defined and the choice variable has a retest correlation of 1, the corresponding correlation for the consequence variable will only in special cases have the value 1.

The attentive reader has from the above already anticipated that the reliability of the sum score of the 27 items (situations) should be great and so it is. Two different types of values are used: the general reliability of a composite measurement and one of its special cases, the so-called Cronbach's alpha coefficient. I have used corrected R^2 values as item reliabilities for the first type of values. The presentation is, for each dependent variable and type of value, in the following order: the decision test, the repetition 1, 2 and 3 of the experiment. The general values are 0.929, 0.901, 0.940 and 0.945 for E and for E_B 0.963, 0.962, 0.979 and 0.978. The alpha coefficients are 0.885, 0.872, 0.904 and 0.899 for E and 0.915, 0.914, 0.938 and 0.927 for E_B. We see the same picture for both types of values: E_B has higher reliability than E and the decision test has almost as high reliability as the experiment. The alpha coefficients are smaller than the general values, which is the normal case, since the alpha coefficient is the general value with average item variance of true score estimated by the average covariance between items. The latter value can never exceed the former value and the alpha coefficient is therefore, according to the classical reliability theory, a conservative measure which can be quite useless if the covariances are small. However,

both types of values may be too high because of a violation of the assumption that measurement errors of items are linearly independent. (I think that this assumption is more realistic for the experiment then for the decision test, as every subject has its own sequence of items.) In the light of this fact, the alpha coefficient may be a more"reliable" reliability measure, since its conservatism may balance the above violation. Anyhow, the reliability of E and E_B is high. Finally, I can mention that the group ANOVAs give reliability values for sum scores of E, E_B and n of 0.941, 0.968 and 0.984, respectively. It can also be mentioned here that the distributions of the sum scores are more regular than the distributions of the single items. The distributions of the sum scores are negatively skewed, but only slightly, with E having somewhat lower means and higher standard deviations than those of E_B.

Other comparisons

The comparisons of the decision test and the experiment are based on means, standard deviations and correlations. For E, the means of the decision test and the three repetitions of the experiment are 0.744, 0.695, 0.715 and 0.731, respectively. The corresponding values for E_B become 0.849, 0.847, 0.847 and 0.852. Remembering that the decision test was given after the experiment, we discover a time trend for E, but not for E_B. However, the differences are small, not greater than 0.050. The average standard deviations for E are 0.165, 0.191, 0.184 and 0.179, while E_B shows 0.130, 0.143, 0.142 and 0.134. Again, we find time trends. Thus, E goes up with time and the group becomes more homogeneous for both E and E_B. The relationships between E and E_B are the same for the decision test as for the experiment. Concerning the correlations, something was already mentioned in connection with reliability. The average correlations between the decision test and the repetitions of the experiment are 0.228, 0.240 and 0.333 for E and 0.305, 0.294 and 0.318 for E_B. The average correlations between repetitions are (0.488, 0.441, 0.598) and (0.613, 0.575, 0.676) for E and E_B, respectively.

Results in accordance with factors D, P and C have already been discussed for the experiment, as far as means and standard deviations are concerned. No ANOVA results have been produced for the

decision test but the different means show the same pattern here as for the experiment, with the possible exception of PC for the dependent variable E. The equivalence is also valid for standard deviations, again with the exception of PC. The decision test shows the same kind of effect for E, although not so pronounced as in the experiment. For E_B, there is no PC effect in the experiment, while the decision test has an effect opposite to that for E: the most asymmetric situations show the smallest standard deviations and the least asymmetric situations show the greatest standard deviations. Why this is so, is hard to say. As we have no corresponding effect for the standard deviations of n, it may show that the symmetric situations are less robust to deviations from the optimal choice of n. This is true for d = 0.2 and 0.3, but not for d = 0.1, and because the situations with d = 0.1 have less effect on E_B variation, due to robustness, it may be generally true.

The correlations show no uniformity at all. There are different effects for the decision test and the experiment as well as for E and E_B and it will not be discussed. The average correlation between the experiment and the decision test is not very great but canonical correlation analyses (between the decision test and each of the repetitions) show that the correlations should not be regarded as unessential. The first canonical correlations are in the neighbourhood of 1.000 with normal deviates of the χ^2 values of 6.9, 7.5 and 5.0 for E and 13.4, 14.1 and 12.2 for E_B. However, the analyses show some numerical instability due to many variables in relation to the number of individuals, and this has also the effect of raising the greatest canonical correlations. I therefore see little reason to discuss these analyses in detail. (A somewhat more reasonable analysis had been to find the correlations with the restriction of equal weight vectors, but no such program was available.)

For both E and E_B, factor analyses have been performed for each of the sets of 27 situations. This kind of factor analysis gives a principal axis solution and a varimax rotation, see Dixon (1967). The communality estimates are squared multiple correlations and only factors with eigenvalues exceeding 1.0 have been rotated. The analysis is not very satisfying, but no program for direct comparisons of structures was available. The dependent variable E gave 6 rotated factors for the deci-

sion test and 5 factors for each repetition. The variable E_B gave 4 rotated factors for the decision test and 5, 4 and 4 for the repetitions of the experiment. The number of factors is reasonable for each analysis, explaining between 0.829 and 0.863 of the total common variance (estimated as the sum of the 27 squared multiple correlations). The average absolute deviation of eigenvalues of successive unrotated factors between the decision test and the repetitions tells us something about the structures. (The sum comprises the five first unrotated factors.) We get 0.37, 0.26 and 0.40 for E, but 1.38, 0.86 and 0.90 for E_B, thus indicating that the distributions of eigenvalues differ more for E_B. Corresponding values between repetitions are 0.44, 0.50 and 0.18 for E and 0.60, 0.52 and 0.28 for E_B, also meaning that the decision test differs more for E_B. Similar results are obtained for factor loadings of the unrotated factors. The average number of loadings (for the first five factors), which differs more than 0.30, when corresponding values of two factor analyses are compared, are 5.9 and 7.1 for E and 10.1 and 6.7 for E_B. The first figure refers to comparisons between the decision test and the experiment and the second one refers to comparisons within the experiment. The first factor shows better equivalence than the others, which are not very similar. The factor analyses seem to show that the decision test is different from the experiment for E_B, relative to the difference within the experiment. No attempts have been made to "interpret" the rotated factors.

Intelligence and efficiency

The intelligence tests show sufficient discriminating ability and have reasonable reliability. (See appendix 3.) Compared to Holmquist's group, my subjects are better when it concerns "factor" V, W and S and worse on "factor" P, where they also are somewhat more homogeneous. The differences are probably due to age differences and to the fact that my university people are a selected group of students leaving the gymnasium. The reliability estimates shown by Holmquist (1967) are, on the average, of the same magnitude as those which are presented in appendix 3 in the column marked with r_1. The estimate r_1 is a special Cronbach's alpha coefficient with the assumption of equal item difficulties, implying that the total mean and variance are

sufficient for estimating the reliability. Since only the total number of correctly answered items was punched for each subject and test, this reliability estimate was the only accessible one. However, it is known to be below the alpha coefficient to an extent which depends on the variance of the item difficulties, see e.g. Horst (1966, p. 273). I have therefore also made estimates on the assumption that the item difficulties are rectangularly distributed, which I believe is more reasonable than the assumption of zero variance. The new estimates are shown in appendix 3 in the column marked with r_2. We see that "factor" S has a somewhat higher average reliability, but almost all estimates r_2 are reasonably high for group comparisons.

Eight canonical correlation analyses have been performed between the ten intelligence tests and the 27 decision situations. For E as well as for E_B the four analyses comprise the decision test and the three repetitions of the experiment. All analyses show the same result. Although the first canonical correlations are about 0.900, their corresponding χ^2 values are not greater than those expected by chance. This is in line with the magnitudes of the correlations between the intelligence tests and the decision situations. The 1,080 correlations for the analyses of E have $-0.299 \leq r \leq 0.486$ and those for E_B have $-0.299 \leq r \leq 0.483$. The S and the I tests have somewhat higher correlations than the other tests, but on the whole these correlation analyses cannot verify any substantial relations between intelligence and efficiency.

Factor analyses have also been performed, of the same kind as was described before. When the intelligence tests are analysed alone we get two rotated factors "explaining" 0.941 of the total common variance. The first one is spatial and inductive, the other one being verbal (V and W). Eight further factor analyses were performed, each comprising the intelligence tests and one efficiency variable, the last one being a sum of 27 efficiency scores. (We have such a sum for the decision test and the three repetitions, partly for E and partly for E_B.) The addition of the efficiency variable does not change the result of the intelligence variables very much. Thus there are always two factors, which "explain" about 90 per cent of the total common variance. The only difference concerns the repetitions, where the

spatial-inductive factor becomes purely spatial.

The communality estimates for the efficiency variable are low with $0.137 \leq R^2 \leq 0.340$, while the corresponding values for the average R^2 show $0.417 \leq \overline{R}^2 \leq 0.444$ for the eight analyses. The only intelligence test, where R^2 is raised when the efficiency variable is added as a tenth independent variable, is one of the spatial test, showing an average increase 0.004 for the decision test and 0.048 for the experiment. The average correlation between these variables are 0.177 and 0.400, respectively. As could be expected, the only substantial loading of the efficiency variable is for the spatial factor, with loadings between 0.242 and 0.465. Thus the result of the canonical correlation analyses reappears: the efficiency of decision making is not very dependent on intelligence, with the possible exception of spatial ability, and this is valid for both E and E_B. I do not even know if it is a purely spatial test: some of those who have used this test assert that it is very frustrating to the subject and also is an endurance test.

spatial-inductive factor becomes purely spatial.

The communality estimates for the efficiency variable are low with $0.137 \leq R^2 \leq 0.340$, while the corresponding values for the average R^2 show $0.417 \leq \bar{R}^2 \leq 0.444$ for the eight analyses. The only intelligence test, where R^2 is raised when the efficiency variable is added as a tenth independent variable, is one of the spatial test, showing an average increase 0.004 for the decision test and 0.048 for the experiment. The average correlation between these variables are 0.177 and 0.400, respectively. As could be expected, the only substantial loading of the efficiency variable is for the spatial factor, with loadings between 0.242 and 0.465. Thus the result of the canonical correlation analyses reappears: the efficiency of decision making is not very dependent on intelligence, with the possible exception of spatial ability, and this is valid for both E and E_B. I do not even know if it is a purely spatial test: some of those who have used this test assert that it is very frustrating to the subject and also is an endurance test.

FINAL COMMENT

Only the experiment is discussed in this section, although some comments are also applicable to the decision test. The items of this test have, on the average, similar values for reliabilities, means and standard deviations as do the situations of the experiment. The correlations between corresponding items and situations are somewhat lower than correlations between situations from repetition to repetition. However, regarded as an instrument which gives an efficiency sum score it is about as good as one repetition of the experiment - and much cheaper. The efficiency seems to be rather unrelated to intelligence as that is defined here, with the possible exception of spatial ability.

The lens model

I thought that this study was a Bayesian study, but after reading Slovic & Lichtenstein (1971) I know better. The content of the study is, of course, a Bayesian one, but, according to their excellent paper, the approach of the study is mainly a regression approach. Brunswik's lens model is more or less applicable to the treatment of data of this report (see appendix 4). The stimulus dimensions or cues are d, c_0/c_1 and $P(H_0)$, which by suitable dummy variable coding give rise to seven effects. The correlations between independent and dependent variables, for a subject or a group of subjects, called utilization coefficients, are squared here and denoted ω^2, which are squared multiple correlations between the dependent variable and the dummy variables defining an effect. Due to orthogonality the squared consistency index r_s^2 is then equal to the sum of the seven ω^2. Low r_s values are said to show inconsistency. You could just as well call it irrelevance, because low r_s values may mean that the subject uses other cues than those which the experimenter thinks he is presenting. Both "explanations" can be more or less correct simultaneously. On the criterion side the corresponding utilization coefficients are called ecological validites, and an index of the environmental predictability r_e is the correspondence of the consistency index. In this study the squared validities are the different ω^2 values of the statistical model and the sum of these values stands for the squared index of the environmental predictability. Neither the achievement index r_a nor the matching index r_m is calculated as the lens model prescribes.

The application of the total lens model is only meaningful for the dependent variable n. Here r_c is 1.0 so that the lens model equation degenerates to $r_a = r_s r_m$, which means that squared achievement index has the sum of the subject's seven ω^2 values as its upper bound. E_B could be regarded as an analogue to r_a, and a better one, because r_a is not sensitive to mean differences of n and is an index for choice variables here. The subject can very well rank n (for the 27 situations) in the same way as the statistician does and still has low efficiency. The opposite may also be true in certain cases: due to small variation of n and robustness we can obtain low r_a values and high E_B values. The dependent variables E and E_B are themselves used in ANOVA, but the lens model is only half here, since the efficiency is always 1.0 for the statistician the criterion side collapses (it is already comprised by the dependent variables). The fourth dependent variable, k_c, cannot, as far as I can see, be used within the scope of the lens model, partly because it depends on n and partly because it is non-numerical when n is zero.

When using the full lens model the statistician is regarded as the criterion with $r_c = 1.0$. This is not necessary, e.g. having the same statistician making sequential observations will produce a r_c value less than 1.0. The criterion can of course also be other things than a statistical model. The most used alternative is "true" data - observations from a follow-up study - but it can just as well be constituted by observations with another response method or the result of another subject or a group of subjects. And there is nothing that prevents you from having a second model on the subject side, thus using the lens model to compare two models. Nor are there any obstacles to generalizing the lens model to a multivariate model, although there will be problems, as for other multivariate models, of creating convenient indices.

The statistical model is used in two ways in this study. For n, it is used on the criterion side of the lens model (and something like that for k_c, too), while for E and E_B the model is used to evaluate the choice variables (to construct the consequence variables). We can say that n is evaluated twice. First by using the lens model to compare the utilization coefficients with the ecological validities (the choice level), second by calculating E_B and looking on its utilization coeffi-

cients (the consequence level). The fact that the full lens model does not work with E_B or E (or the often used accuracy ratio) is not a general property for a consequence variable. For instance, it will work with R_B and R. However, the full lens model will, for some cases, collapse when the criterion side is occupied by the same entity as that which is used for the construction of the consequence variable. Its values for this entity are then the same constant for all situations. This will in general not happen if one uses different entities for the two purposes, e.g. if the criterion side is represented by a special subject and the statistical model is only used to get E and E_B.

Results

The discussion here builds on the paper by Slovic & Lichtenstein (1971). This is hardly any restriction, since this paper broadly reviews much research in the Bayesian area and other kind of research using the lens model. Although the paper is almost only concerned with probability (revision) experiments - and not with information-seeking experiments, which this study is - some concepts and results can still be applied and discussed here, at least in connection with the choice variables k_c and n.

One of the key concepts in Bayesian research is that of conservatism. Apparently this word means different things to different researchers. For the one who performs a probability revision experiment it means that the subject makes too small a revision in comparison with the prescription of Bayes's theorem, and this can be measured in several ways. Others have used $\sqrt{1 - r_s^2}$ or compared the utilization coefficients with the ecological validities. The crucial issue is whether you will define conservatism as a measure of distance or as a measure of (co)variance. Take factor D for the dependent variable n as an example. The statistician has means 7.4, 29.7 and 24.8 for increasing d values, while the corresponding means for the total group of subjects are 23.2, 20.6 and 19.5. For the statistician ω_D^2 is 0.676 and for the average subject it is 0.134. The distance measure shows that this subject is conservative for d = 0.1 and radical otherwise. The variance of the means of the subject is less than that of the statistician, so from this point of view the subject is conservative. Comparing ω^2 values will also result in conservatism here. As you can see no choice can tell the whole story and different indices can

classify subjects differently. My choice is to define the degree of conservatism as ω^2 (statistician)/ω^2 (subject), for a certain effect. This implies that conservatism is defined as a lack of relative variance. If you like it, you may also say that conservatism, for an effect, means lesser diagnosticity than the model prescribes. The above ratio should only be of interest when ω^2 (subject) is sufficiently high. Although a ratio of 10 indicates a considerable conservatism, I hesitate to find it essential if we have e.g. 10 = 0.020/0.002.

In trying to explain conservatism one has used the labels misperception, misaggregation and response bias. Misperception means subjective transformation of the cue values, misaggregation means that the subject's policy for using the cues in order to generate a value of the dependent variable is deviant from the model while response bias can involve such things as sensibility to different response modes and the range of the cue values. It is not often that experiments are designed to differentiate between these explanations, and like any other information-seeking experiments which I know of this study cannot differentiate between the possible explanations. Of course, this does not prevent you from discussing them.

Although trivial, it is perhaps best to underline that conservatism as well as its explanations are relative concepts. While a subject may be conservative versus one model (or another subject) he may be radical in comparison with another model (or a third subject), and while one model classifies your judgments as misperceptions, another one may call them misaggregations, or both. I think that for every consistent behaviour you can construct a model which, on the average, describes this behaviour. This is not very interesting as it presumably means one model for each subject. However, the models and thereby the subjects can be clustered according to certain properties to obtain more general knowledge. (Analogous ways have been tried, which in this case could have involved a data matrix of order 60 x 7 with the seven ω^2 values as variables. Some kind of method for latent structure analysis could then be used to cluster the subjects into subgroups of similar ω^2 profiles.) Instead of doing this very extensive labour the researcher chooses prior models with which he compares the subjective behaviour. Different camps of researcher have different such models and therefore can have different explanations of "deviant"

behaviour. It may thus be wise not to say e.g. misaggregation but misaggregation in comparison with a Bayes strategy.

The ecological validities for n are small for four effects, very high for D and noticeable for PC and DPC. According to our definition, all subjects show conservatism for D and all but one for DPC, too. On the other hand, several subjects show radicalism, especially for P but we may also mention PC. This is not in line with Slovic & Lichtenstein (1971), who report that interaction effects have small increments in predictive power. Here, 17 out of 60 subjects have results such that $0.100 < \omega^2_{PC} \leq 1.000$. It is also said that the most important cue usually accounts for more than 40 per cent of the predictable variance ($\max(\omega^2)/\Sigma\omega^2$) and I can somewhat agree with it. Twenty subjects show this result, seven allocated on D, seven on P, one on C and five on PC. However, the statement that the three most important cues usually cover more than 80 per cent cannot be confirmed. This is only valid for five subjects and the statistician. Thus, the majority of the subjects is not focusing on a single cue and they have quite varying squared consistency indices, $0.160 \leq r_s^2 \leq 1.000$ with a mean of 0.596, with very different ω^2 profiles.

The most remarkable feature about the choice of k_c is the asymmetry. Most subjects do not "like" H_0, at least for large n, and more or less consequently choose k_c less than k_B, the most extreme choice being $k_c = 10$ for $n = 100$. This implies that there is a tendency for several subjects to choose k_c more extreme than k_B for situations with $k_B < n/2$. This does not seem to be in line with the mainstream of results either. Slovic & Lichtenstein (1971) say that subjects are never as sensitive to the experimental conditions as they ought to be for the Bayesian research they have summarized. However, the extent to which this statement does not hold in this study is dependent on the choice of a criterion. For instance, there are 45 subjects with $B < 0.5$ ($k_c = A + Bn$), but there are only 20 subjects having $s^2(k_c)/s^2(n)$ greater than the corresponding ratio for the model.

I have earlier in this paper suggested that the asymmetry of k_c may be caused by an asymmetry of the apprehension of 0 and 1. This gives rise to a misperception of the binomial frequency function, which has been experimentally verified before. If the outcome 1 has a greater impact than the outcome 0 we will get $B < 0.5$, provided no misaggrega-

tions occur, that is, the subject uses the Bayes strategy within his subjective apprehension (misperception) of the binomial frequency function. Another alternative for misperception: suppose that the subject does not quite trust the data. This can generate reliability models like those in Shum & DuCharme (1971), which perhaps can be used to "emplain" the asymmetry of k_c. I also mentioned earlier (p. 22) that the response mode can have caused the bias.

The above examples of misperception may also be used as a descriptive model for some subjects and perhaps also describe some subjects' choices of both k_c and n. But, in comparison with the Bayesian model, subjects most likely also misaggregate cues when choosing k_c and n. (Provided no misperception, this is e.g. reflected by the ω^2 profile.) The situations are complex and I think that the subjects simplify reality by creating simple rules. These can be followed more or less rigorously. A few subjects have rules, which I can see have been followed all the way, e.g. "I never make any observations" and "Regardless of d, I make 10 observations when $P(H_0)$ and c_0/c_1, balance each other (e.g. H_0 improbable but cheap) and make no observations in other cases". Examples of rules which are almost always followed are "If c_0/c_1, is 1.0 I will make 10 observations, otherwise I make 20 observations," and "If $P(H_0)$ is 0.3 I will make 20 observations, otherwise I make 10 observations". Then there may be more stochastic rules like "I always make between 10 and 40 observations, but for every trial I just guess".

This scattered picture can make you rather pessimistic about ever finding any general results. It is quite clear that it is very dangerous to present only group results. As the individual strategies vary considerably you can get almost arbitrarily varying group results by changing the group composition. We may also remember that group results can give peculiar results in comparison to the individual results, due to lack of commutability. For instance, ω^2 for the average subject derived from the group ANOVA is not equal to the average ω^2 calculated from the individual ANOVAs. Not knowing such properties can generate more or less unreasonable conclusions. It is easily done, because I believe that most of us try to look upon the world as simply as possible, perceive situations as symmetric, commutative, full of linear relations and so on.

One bold solution to this multitude of behaviour is to neglect it. For instance, who cares about conservatism or radicalism if the efficiency of the resulting decision is high? Although this is an extreme opinion there is a kernel of truth in it. It is somewhat strange that, while decision theory itself preaches that it is the consequences which count, researchers on human decision making usually concentrate on choice variables. I think we can add another dimension to the discussion of deviant behaviour if we also consider its consequences when possible. Psychologists naturally have an interest in choice variables, but from an "economic" viewpoint these will only be of importance for situations where they indicate non-optimal behaviour of low efficiency.

The consequence variables E and E_B have, on the average, higher consistency than the choice variable n. We obtain $0.291 \leq r_s^2 \leq 0.979$ with a mean of 0.718 for E and $0.304 \leq r_s^2 \leq 1.000$ with a mean of 0.773 for E_B. Although E_B has a higher mean than E, it is not so for every subject. We also have $r_s^2(E_B) = 1.000$ if $r_s^2(n) = 1.000$ or when the subject always makes the same number of observations (in which case $r_s^2(n)$ is not defined). However, I do not know whether, for two subjects i and j, $r_s^2(n_i) < r_s^2(n_j)$ implies $r_s^2(E_{Bi}) < r_s^2(E_{Bj})$. Probably not. As for n, the ω^2 profiles for E and E_B are very different from subject to subject, but factor I has the largest average utilization coefficients and only DPC for E_B has also an average ω^2 above 0.100. Speaking about consequences, D is the most important cue for most subjects, but its different levels are not of the same interest. It is, above all, d = 0.3 which tests the subjects, while d = 0.1 has low discriminating power (or high robustness). The latter situation is analogous to dealing with intelligent persons: no matter how you teach them, some elementary material will they learn.

I do not know how common situations, as those with d = 0.1 are, but I believe that a great many situations can be described by criterion functions which are flat around its optimal point. On the other hand, there are definitely situations where choices are crucial. My proposal is that more experiments should be designed with the latter kind of situations. This may not be easy, but it will add an importance to the choice variables they often not have today.

REFERENCES

Bradley, J. V. Distribution-free statistical tests. Englewood Cliffs, New Jersey: Prentice Hall, 1968.

Darlington, R. B. Multiple regression in psychological research and practice. (1968) In Lieberman, B. (Ed.) Contemporary problems in statistics. New York: Oxford Univer. Pr., 1971. Pp. 384-407.

DeGroot, M. H. Optimal statistical decisions. New York: McGraw-Hill, 1970.

Dempster, A. P. Elements of continuous multivariate analysis. Reading, Mass., Addison-Wesley, 1969.

Dixon, W. J. (Ed.) BMD. Biomedical computer programs. Berkeley, Calif.: Univer. Calif. Pr., 1967.

Dixon, W. J. (Ed.) BMD. Biomedical computer programs. X-series supplement. Berkeley, Calif.: Univer. Calif. Pr., 1970.

Hays, W. & Winkler, R. Statistics: Probability, inference and decision. New York: Holt, 1971.

Holmquist, R. The test battery for matric candidates. Meddelande från Centraltestoteket (PA-rådet), nr 15, 1967. (In Swedish)

Horst, P. Psychological measurement and prediction. Belmont, Calif.: Wadsworth, 1966.

Kogan, N. & Wallach, M. A. Risk taking. A study in cognition and personality. New York: Holt, 1964.

Larsson, B. Bayes strategies and human information seeking. Lund: Gleerup, 1968.

Larsson, B. Efficiency of some Bayesian decision procedures. Didakometry, Nr 26, 1970.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Pratt, J. W., Raiffa, H. & Schlaifer, R. Introduction to statistical decision theory. New York: McGraw-Hill, 1965.

Shum, D. A. & Du Charme, W. M. Comments on the relationship between the impact and the reliability of evidence. Organiz. Behav. Human Perform, 1971, 6, 111-131.

Slovic, P. & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organiz. Behav. Human Perform, 1971, 6, 649-744.

Snapper, K. J. & Peterson, C. Information seeking and data diagnosticity. Stencil, 1971.

Staël von Holstein, C.-A. S. Assessment and evaluation of subjective probability distributions. Stockholm: Economic Research Institute, Stockholm, School of Economics, 1970.

APPENDICES

Appendix 1. Choices and expected losses of the statistical model.

$P(H_0)$	c_0^d/c_1	0.1			0.2			0.3		
		0.5	1.0	2.0	0.5	1.0	2.0	0.5	1.0	2.0
0.3	n_o	20	0	0	31	31	23	25	24	23
	k_B	9	H_1	H_1	15	14	9	12	11	10
	R_o	1039	900	600	644	650	594	407	414	395
0.5	n_o	0	27	0	31	35	31	26	27	26
	k_B	H_0	13	H_1	16	17	14	13	13	12
	R_o	1000	1169	1000	665	693	665	419	431	419
0.7	n_o	0	0	20	23	31	31	23	24	25
	k_B	H_0	H_0	10	13	16	15	12	12	12
	R_o	600	900	1039	594	650	644	395	414	407

When $n_o = 0$ the hypothesis chosen is indicated for k_B. The unit of R_o is one Swedish öre, which for this experiment constitutes one tenth of the cost of one observation.

Appendix 2. The distribution of subjects on faculty, sex and age.

Faculty of	Sex	Age 19-24	25-29	30-45
Humanities	Female	7	4	1
	Male	2	1	1
Social sciences	Female	7	8	0
	Male	7	6	5
Natural sciences	Female	1	0	0
	Male	7	3	0

Humanities: 16

Social sciences: 33

Natural sciences: 11

Female: 28

Male: 32

19-24: 31

25-29: 22

30-45: 7

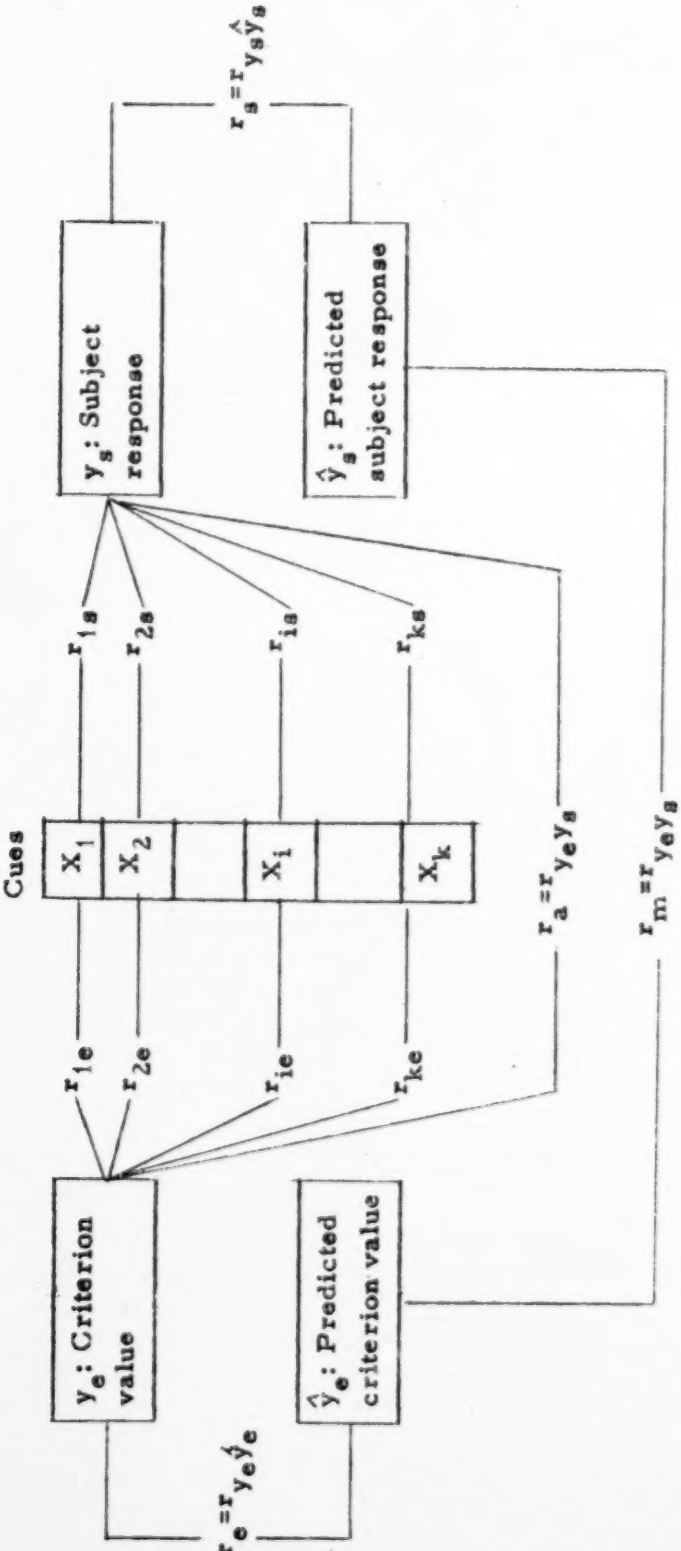
Appendix 3. The intelligence tests.

Test	m	s	r_1	r_2
V_1	29.6	6.4	0.771	0.864
V_2	21.7	3.1	0.585	0.825
W_1	33.8	8.2	0.855	0.918
W_2	19.6	5.7	0.648	0.778
I_1	17.1	4.3	0.733	0.852
I_2	16.3	5.1	0.746	0.844
S_1	35.9	13.4	0.935	0.963
S_2	24.0	6.2	0.767	0.856
P_1	35.3	6.3	0.623	0.757
P_2	22.1	8.0	0.886	0.935

The tests V_1 and V_2 concern verbal understanding, one is a test on synonymns and the other contains items on verbal analogies. The next two tests, W_1 and W_2, are tests on verbal fluency. The task of W_1 is to write as many words as possible, which begin with "s" and end with "a", while the task of W_2 concerns words which end with "al". The test I_1 and I_2 measure inductive reasoning and the items of both tests are series of numbers for which a new number should be added. S_1 and S_2 are spatial tests, the items of which are three-dimensional bodies unfolded in two dimensions and the task is to say something about their three-dimensional forms. The final tests, P_1 and P_2, are supposed to measure the perceptual factor. One of them has to do with sorting and the other concerns coding.

Regarding the columns of the above table, m stands for the arithmetic mean, s is the standard deviation, r_1 is a specialized Cronbach's alpha coefficient (also known as Kuder-Richardson's formula 21) and r_2 is a special estimate of Cronbach's alpha coefficient as it is discussed on page 35.

Appendix 4. The lens model (After Slovic & Lichtenstein, 1971).



The lens model equation: $r_a = r_e r_s r_m + c((1-r_e^2)(1-r_s^2))^{1/2}$

r_{is}: Utilization coefficient for cue x_i,

r_{ie}: Ecological validity for cue x_i.

r_s: Consistency index.

r_e: Index of environmental predictability.

r_a: Achievement index.

r_m: Matching index.

c: Defined as $r_{(y_e-\hat{y}_e)(y_s-\hat{y}_s)}$.

Both $\hat{y}_s$ and $\hat{y}_e$ are linear regression functions of the cue values.

For ANOVA, r_a is equal to the total correlation, r_m is a between-cells correlation and c is a within-cells correlation, while r_s and r_e are ω between cells, as described in the Final comment.

Appendix 5. Symbols used frequently.

A	The intercept of $k_c = A + Bn$.
B	The regression coefficient of $k_c = A + Bn$.
C	Factor of cost ratios c_0/c_1 with levels 0.5, 1.0 and 2.0.
c_0	The loss generated by a wrong choice of H_0.
c_1	The loss generated by a wrong choice of H_1.
D	Factor of d values with levels 0.1, 0.2 and 0.3.
d	Diagnosticity of data, defined as $d = p_1 - p_0$.
E	The efficiency of the choices of k_c and n, defined as $E = R(k_B, n_o)/R(k_c, n)$ or, shorter, R_o/R.
E_B	The efficiency of the choice of n, defined as $E_B = R(k_B, n_o)/R(k_B, n)$ or, shorter, R_o/R_B.
H_0	The null hypothesis $p = p_0$
H_1	The alternative hypothesis $p = p_1$.
k	The number of ones of n observations.
k_B	The critical value of k according to the statistical model. It chooses H_0 if $k \leq k_B$ and H_1 otherwise. (The k value of the Bayes strategy.)
k_c	The critical value of k chosen by a subject. He chooses H_0 if $k \leq k_c$ and H_1 otherwise.
m	The arithmetic mean.
n	The number of observations (chosen by a subject).
n_{o_2}	The number of observations according to the statistical model.
ω	Hays' ω^2, defined as SS_i/SS_{total} for an effect i. It is a squared multiple correlation between the dependent variable and the dummy coded variables defining the effect.
P	Factor of prior probabilities $P(H_0)$ with levels 0.3, 0.5 and 0.7.
P(.)	The probability of something. Especially, $P(H_0)$ is the prior probability of H_0 (before sampling) and $P(H_0 \mid k, n)$ is the posterior probability of H_0 (after sampling, when k and n are known).
p_0	The probability that an observation will have the outcome 1 (according to H_0).
p_1	The probability that an observation will have the outcome 1 (according to H_1).
R^2	The squared multiple correlation.

$R(k_B, n) = R_B$ The total expected loss of choosing n observations and using the critical value k_B.

$R(k_c, n) = R$ The total expected loss of choosing n observations and using the critical value k_c.

$R(k_B, n_o) = R_o$ The total expected loss of choosing n_o observations and using the critical value k_B.

r The product-moment correlation.

S Factor of the subjects with 60 levels.

SS Sum of squares.

s The standard deviation.

T Factor of replications with three levels.